Averaging over fast variables in the fluid limit for Markov chains: application to the supermarket model with memory

M.J. Luczak [1] [2] and J.R. Norris [3] [4]

January 7, 2010

Abstract

We set out a general procedure which allows the approximation of certain Markov chains by the solutions of differential equations. The chains considered have some components which oscillate rapidly and randomly, while others are close to deterministic. The limiting dynamics are obtained by averaging the drift of the latter with respect to a local equilibrium distribution of the former. Some general estimates are proved under a uniform mixing condition on the fast variable which give explicit error probabilities for the fluid approximation.

Mitzenmacher, Prabhakar and Shah [8] introduced a variant with memory of the 'join the shortest queue' or 'supermarket' model, and obtained a limit picture for the case of a stable system in which the number of queues and the total arrival rate are large. In this limit, the empirical distribution of queue sizes satisfies a differential equation, while the memory of the system oscillates rapidly and randomly. We illustrate our general fluid limit estimate in giving a proof of this limit picture.

Keywords. Join the shortest queue, supermarket model, supermarket model with memory, law of large numbers, exponential martingale inequalities, fast variables, correctors.

AMS (2000) Mathematics Subject Classification. Primary 60J28; Secondary 60K25

1 A general fluid limit estimate

We describe a general framework to allow the incorporation of averaging over fast variables into fluid limit estimates for Markov chains, building on the approach used in [2]. The

[1]Department of Mathematics, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: m.j.luczak@lse.ac.uk

[2]Supported in part by a STICERD grant. Part of this work was accomplished while visiting the Mittag-Leffler Institute.

[3]University of Cambridge, Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB, UK. Email: j.r.norris@statslab.cam.ac.uk

[4]Research supported by EPSRC grant EP/E01772X/1.

main results of this section, Theorems 1.5 and 1.6, establish explicit error probabilities for the fluid approximation under assumptions which can be verified from knowledge of the transition rates of the Markov chain. See also [1] for related results.

1.1 Outline of the method

Let $X = (X_t)_{t\geqslant 0}$ be a continuous-time Markov chain with countable state-space S and with generator matrix $Q = (q(\xi, \xi') : \xi, \xi' \in S)$. Assume that the total jump rate $q(\xi)$ is finite for all states ξ, and that X is non-explosive. Then the law of X is determined uniquely by Q and the law of X_0. Make a choice of *fluid coordinates* $x^i : S \to \mathbb{R}$, for $i = 1, \dots, d$, and write $\boldsymbol{x} = (x^1, \dots, x^d) : S \to \mathbb{R}^d$. Consider the $\mathbb{R}^d$-valued process $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geqslant 0}$ given by $\boldsymbol{X}_t = (X_t^1, \dots, X_t^d) = \boldsymbol{x}(X_t)$. Call $\boldsymbol{X}$ the *slow* or *fluid variable*. Define for each $\xi \in S$ the *drift vector*

$$\beta(\xi) = Q\boldsymbol{x}(\xi) = \sum_{\xi'\neq\xi}(\boldsymbol{x}(\xi') - \boldsymbol{x}(\xi))q(\xi, \xi').$$

Make also a choice of *auxiliary coordinate* $y : S \to I$, for some countable set I, and set $Y_t = y(X_t)$. Call the process $Y = (Y_t)_{t\geqslant 0}$ the *fast variable*. For $\xi \in S$ and $y' \in I$ with $y' \neq y(\xi)$, write $\gamma(\xi, y')$ for the total rate at which Y jumps to y' when X is at ξ. Thus

$$\gamma(\xi, y') = \sum_{\xi':y(\xi')=y'} q(\xi, \xi').$$

Choose a subset U of $\mathbb{R}^d$ and a function $b : U \times I \to \mathbb{R}^d$. Choose also, for each $x \in U$, a generator matrix $G_x = (g(x, y, y') : y, y' \in I)$ having a unique invariant distribution $\pi_x = (\pi(x, y) : y \in I)$. These choices are to be made so that $\beta(\xi)$ is close to $b(\boldsymbol{x}(\xi), y(\xi))$ and $\gamma(\xi, y')$ is close to $g(\boldsymbol{x}(\xi), y(\xi), y')$ whenever $\boldsymbol{x}(\xi) \in U$ and $y' \in I$. Define for $x \in U$

$$\bar{b}(x) = \sum_{y\in I} b(x, y)\pi(x, y).$$

Then, under regularity assumptions to be specified later, there exists a function $\chi : U \times I \to \mathbb{R}^d$ such that

$$G\chi(x, y) = \sum_{y'\in I} g(x, y, y')\chi(x, y') = b(x, y) - \bar{b}(x). \tag{1}$$

Make a choice of such a function χ. Call χ the *corrector for b*.

Fix $x_0 \in U$. We will assume that $\bar{b}$ is Lipschitz on U. Then the differential equation $\dot{x}_t = \bar{b}(x_t)$ has a unique maximal solution $(x_t)_{t<\zeta}$ in U starting from x_0. Fix $t_0 \in [0, \zeta)$. Then for $t \leqslant t_0$

$$x_t = x_0 + \int_0^t \bar{b}(x_s)ds. \tag{2}$$

Define for $\xi \in S$ with $\boldsymbol{x}(\xi) \in U$

$$\bar{\boldsymbol{x}}(\xi) = \boldsymbol{x}(\xi) - \chi(\boldsymbol{x}(\xi), y(\xi)).$$

Let T a stopping time such that $\boldsymbol{X}_t \in U$ for all $t \leqslant T$. Then, under regularity assumptions to be specified later, for $t \leqslant T$

$$\bar{\boldsymbol{x}}(X_t) = \bar{\boldsymbol{x}}(X_0) + M_t + \int_0^t \bar{\beta}(X_s)ds \tag{3}$$

where $M = M^{\bar{\boldsymbol{x}}}$ is a martingale and where

$$\bar{\beta} = Q\bar{\boldsymbol{x}} = \beta - Q(\chi(\boldsymbol{x}, y)). \tag{4}$$

On subtracting equations (2) and (3) we obtain for $t \leqslant T \wedge t_0$

$$\begin{aligned} \boldsymbol{X}_t - x_t = \boldsymbol{X}_0 - x_0 + \chi(\boldsymbol{X}_t, Y_t) - \chi(\boldsymbol{X}_0, Y_0) + M_t + \int_0^t \Delta(X_s)ds \\ + \int_0^t (\beta(X_s) - b(\boldsymbol{X}_s, Y_s))ds + \int_0^t (\bar{b}(\boldsymbol{X}_s) - \bar{b}(x_s))ds \end{aligned} \tag{5}$$

where $\Delta = G\chi(\boldsymbol{x}, y) - Q(\chi(\boldsymbol{x}, y))$.

The discusson in the present paragraph is intended for orientation only, and will play no essential role in the derivation of our results. Fix $U_0 \subseteq U$ such that for all $\xi, \xi' \in S$ with $\boldsymbol{x}(\xi) \in U_0$ and $q(\xi, \xi') > 0$ we have $\boldsymbol{x}(\xi') \in U$. Assume that T is chosen so that $\boldsymbol{X}_t \in U_0$ for all $t \leqslant T$. Define for $\xi \in S$ with $\boldsymbol{x}(\xi) \in U_0$ the *diffusivity tensor* $\alpha(\xi) \in \mathbb{R}^d \otimes \mathbb{R}^d$ by

$$\alpha^{ij}(\xi) = \sum_{\xi' \neq \xi} (\bar{\boldsymbol{x}}^i(\xi') - \bar{\boldsymbol{x}}^i(\xi))(\bar{\boldsymbol{x}}^j(\xi') - \bar{\boldsymbol{x}}^j(\xi))q(\xi, \xi') \tag{6}$$

and define for $t \leqslant T$

$$N_t = M_t \otimes M_t - \int_0^t \alpha(X_s)ds.$$

Then, under regularity assumptions, N is a martingale in $\mathbb{R}^d \otimes \mathbb{R}^d$. Choose a function $a : U_0 \times I \to \mathbb{R}^d \otimes \mathbb{R}^d$ and set

$$\bar{a}(x) = \sum_{y \in I} a(x, y)\pi(x, y).$$

This choice is to be made so that $\alpha(\xi)$ is close to $a(\boldsymbol{x}(\xi), y(\xi))$ whenever $\boldsymbol{x}(\xi) \in U_0$. Suppose we can find also a corrector for a, that is, a function $\tilde{\chi} : U_0 \times I \to \mathbb{R}^d \otimes \mathbb{R}^d$ such that

$$G\tilde{\chi}(x, y) = a(x, y) - \bar{a}(x). \tag{7}$$

Then, for $t \leqslant T$,

$$\int_0^t \alpha(X_s)ds = \tilde{\chi}(\boldsymbol{X}_t, Y_t) - \tilde{\chi}(\boldsymbol{X}_0, Y_0) - \tilde{M}_t + \int_0^t \tilde{\Delta}(X_s)ds$$
$$+ \int_0^t (\alpha(X_s) - a(\boldsymbol{X}_s, Y_s))ds + \int_0^t \bar{a}(\boldsymbol{X}_s)ds \tag{8}$$

where $\tilde{\Delta} = G\tilde{\chi}(\boldsymbol{x}, y) - Q(\tilde{\chi}(\boldsymbol{x}, y))$ and, under suitable regularity conditions, $\tilde{M} = M^{\tilde{\chi}}$ is a martingale up to T.

The martingale terms M and $\tilde{M}$ in (5) and (8) can be shown to be small, under suitable conditions, using the following standard type of exponential martingale inequality. In the form given here it may be deduced, for example, from [2, Proposition 8.8] by setting $f = \theta\phi$, $A = \theta^2 e^{\theta J}\varepsilon/2$ and $B = \theta\delta$.

Proposition 1.1. *Let ϕ be a function on S. Define*

$$M_t = M_t^{\phi} = \phi(X_t) - \phi(X_0) - \int_0^t Q\phi(X_s)ds.$$

Write $J = J(\phi)$ for the maximum possible jump in $\phi(X)$, thus

$$J = \sup_{\xi, \xi' \in S,\, q(\xi, \xi') > 0} |\phi(\xi') - \phi(\xi)|.$$

Define a function $\alpha = \alpha^{\phi}$ on S by

$$\alpha(\xi) = \sum_{\xi' \neq \xi} \{\phi(\xi') - \phi(\xi)\}^2 q(\xi, \xi').$$

Then, for all $\delta, \varepsilon \in (0, \infty)$ and all stopping times T, we have

$$\mathbb{P}\left(\sup_{t \leqslant T} M_t \geqslant \delta \text{ and } \int_0^T \alpha(X_t)dt \leqslant \varepsilon\right) \leqslant \exp\left\{-\delta^2/(2\varepsilon e^{\theta J})\right\}$$

where $\theta \in (0, \infty)$ is determined by $\theta e^{\theta J} = \delta/\varepsilon$.

Now, if β, γ, α are well approximated by b, g, a and if we can show that the corrector terms in (5) and (8) are insignificant, then we may hope to use these equations to show that the path $(x_t : t \leqslant t_0)$ provides a good (first order) approximation to $(\boldsymbol{X}_t : t \leqslant t_0)$ and moreover that the fluctuation process $(\boldsymbol{X}_t - x_t : t \leqslant t_0)$ is approximated (to second order) by a Gaussian process $(F_t : t \leqslant t_0)$ given by

$$F_t = F_0 + B_t + \int_0^t \nabla\bar{b}(x_s).F_s ds$$

where $(B_t : t \leqslant t_0)$ is a zero-mean Gaussian process in $\mathbb{R}^d$ with covariance

$$\mathbb{E}(B_s \otimes B_t) = \int_0^{s \wedge t} \bar{a}(x_r) dr.$$

Our aim in the rest of this section is to give an explicit form of the first order approximation with optimal error scale, that is, of the same order as the scale of deviation predicted by the second order approximation. The next subsection contains some preparatory material on correctors. A reader who wishes to understand only the statement of the fluid limit estimate can skip directly to Subsection 1.3.

1.2 Correctors

In order to implement the method just outlined, it is necessary either to come up with explicit correctors or to appeal to a general result which guarantees the existence, subject to verifiable conditions, of correctors with good properties. In this subsection we obtain such a general result. In fact, we shall find conditions which guaranteee the existence, for each bounded measurable function f on $U \times I$, of a good *corrector for* f, that is to say a function $\chi = \chi_f$ on $U \times I$ such that

$$G\chi(x, y) = f(x, y) - \bar{f}(x)$$

where

$$\bar{f}(x) = \sum_{y \in I} f(x, y)\pi(x, y).$$

Moreover we shall see that χ_f depends linearly on f and we shall obtain a uniform bound and a continuity estimate for χ_f.

Assume that there is a constant $\nu \in (0, \infty)$ such that, for all $x \in U$ and all $y \in I$, the total rate of jumping from y under G_x does not exceed ν. Then we can choose an auxiliary measurable space E, a family of probability measures $\mu = (\mu_x : x \in U)$ on E, and a measurable function $F : I \times E \to I$ such that, for all $x \in U$ and all $y, y' \in I$ distinct,

$$g(x, y, y') = \nu\mu_x(\{v \in E : F(y, v) = y'\}). \tag{9}$$

Let $N = (N(t) : t \geqslant 0)$ be a Poisson process of rate ν. Fix $x \in U$ and let $V = (V_n : n \in \mathbb{N})$ be a sequence of independent random variables in E, all with law μ_x. Thus

$$g(x, y, y') = \nu\mathbb{P}(F(y, V_n) = y')$$

for all pairs of distinct states y, y' and all n. Fix a reference state $\bar{y} \in I$. Given $y \in I$, set $Z_0 = y$ and $\bar{Z}_0 = \bar{y}$ and define recursively for $n \geqslant 0$

$$Z_{n+1} = F(Z_n, V_{n+1}), \quad \bar{Z}_{n+1} = F(\bar{Z}_n, V_{n+1}).$$

Set $Y_t = Z_{N(t)}$ and $\bar{Y}_t = \bar{Z}_{N(t)}$. Then $Y = (Y_t)_{t \geqslant 0}$ and $\bar{Y} = (\bar{Y}_t)_{t \geqslant 0}$ are both Markov chains in I with generator matrix G_x, starting from y and $\bar{y}$ respectively[5], and are realized on the same probability space. We call the triple (ν, μ, F) a *coupling mechanism*. Define the *coupling time*

$$T_c = \inf\{t \geqslant 0 : Y_t = \bar{Y}_t\}.$$

Assume that, for some positive constant τ, for all $x \in U$ and all $y, \bar{y} \in I$,

$$m(x, y, \bar{y}) = \mathbb{E}_{(x,y,\bar{y})}(T_c) \leqslant \tau. \tag{10}$$

Fix a bounded measurable function f on $U \times I$ and set

$$\chi(x, y) = \mathbb{E}_{(x,y)} \int_0^{T_c} (f(x, Y_t) - f(x, \bar{Y}_t))dt. \tag{11}$$

Then χ is well defined and, for all $x \in U$ and all $y \in I$,

$$|\chi(x, y)| \leqslant 2\tau \|f\|_\infty. \tag{12}$$

Proposition 1.2. *The function χ is a corrector for f.*

Proof. We suppress in the proof the variable x. Note first that if instead of taking $\bar{Z}_0 = \bar{y}$ we start $\bar{Z}$ randomly, with the invariant distribution π, then we change the value of χ by a constant independent of y. Hence it will suffice to establish the corrector equation $G\chi = f - \bar{f}$ in this case. Fix $\lambda > 0$ and define

$$\phi^\lambda(y) = \mathbb{E} \int_0^{T_\lambda} f(Y_t)dt, \quad \bar{\phi}^\lambda = \mathbb{E} \int_0^{T_\lambda} f(\bar{Y}_t)dt$$

where $T_\lambda = T_1/\lambda$, with T_1 an independent exponential random variable of parameter 1. Then, since Y and $\bar{Y}$ coincide after T_c,

$$\bar{\phi}^\lambda - \phi^\lambda(y) = \mathbb{E} \int_0^{T_\lambda \wedge T_c} (f(\bar{Y}_t) - f(Y_t))dt \to \chi(y)$$

as $\lambda \to 0$. By elementary conditioning arguments, $(G - \lambda)\phi^\lambda + f = 0$ and $\lambda \bar{\phi}^\lambda = \bar{f}$, so

$$(G - \lambda)(\bar{\phi}^\lambda - \phi^\lambda) = f - \bar{f}.$$

On passing to the limit $\lambda \to 0$ in this equation, using bounded convergence, we find that $G\chi = f - \bar{f}$, as required. □

[5]The process Y introduced here is not the fast variable, also denoted Y in the rest of the paper: the current Y is to be considered as a local approximation of the fast variable.

We remark that the corrector $\chi(x,.)$ in fact depends only on f, G_x and the choice of $\bar{y}$, as the preceding proof makes clear. The further choice of a coupling mechanism is a way to obtain estimates on χ.

The following estimate will be used in dealing with the Δ term in equation (5). Write $\|\mu_x - \mu_{x'}\|$ for the total variation distance between μ_x and $\mu_{x'}$.

Proposition 1.3. *For all $x, x' \in U$ and all $y \in I$,*

$$|\chi(x,y) - \chi(x',y)| \leqslant 2\tau \sup_{z\in I} |f(x,z) - f(x',z)| + 2\nu\tau^2 \|f\|_\infty \|\mu_x - \mu_{x'}\|. \tag{13}$$

Proof. By a standard construction there exists a sequence of independent random variables $((V_n, V_n') : n \in \mathbb{N})$ in $E \times E$ such that V_n has distribution μ_x, V_n' has distribution $\mu_{x'}$ and $\mathbb{P}(V_n \neq V_n') = \frac{1}{2}\|\mu_x - \mu_{x'}\|$, for all n. Write $(\mathcal{F}_t)_{t\geqslant 0}$ for the filtration of the marked Poisson process obtained by marking N with the random variables (V_n, V_n'). Construct $(Y, \bar{Y})$ from N and $(V_n : n \in \mathbb{N})$ as above. Similarly construct $(Y', \bar{Y}')$ from N and $(V_n' : n \in \mathbb{N})$. Recall that $T_c = \inf\{t \geqslant 0 : Y_t = \bar{Y}_t\}$ and set $T_c' = \inf\{t \geqslant 0 : Y_t' = \bar{Y}_t'\}$. Set $\lambda = \frac{1}{2}\nu\|\mu_x - \mu_{x'}\|$ and set

$$D = \inf\{t \geqslant 0 : (Y_t, \bar{Y}_t) \neq (Y_t', \bar{Y}_t')\}.$$

Then the process $t \mapsto 1_{\{D\leqslant t\}} - \lambda t$ is an $(\mathcal{F}_t)_{t\geqslant 0}$-supermartingale and T_c is an $(\mathcal{F}_t)_{t\geqslant 0}$-stopping time. So, by optional stopping, we have $\mathbb{P}(D \leqslant T_c) \leqslant \lambda \mathbb{E}(T_c) \leqslant \lambda\tau$. Moreover, by the strong Markov property, on $\{D \leqslant T_c\}$, we have $\mathbb{E}(T_c - D|\mathcal{F}_D) = m(x, Y_D, \bar{Y}_D) \leqslant \tau$ so, for any function $g : I \to \mathbb{R}^d$, with $|g| \leqslant \|f\|_\infty$,

$$\mathbb{E}\left|\int_{D\wedge T_c}^{T_c} g(Y_t)dt\right| \leqslant \tau\|f\|_\infty \mathbb{P}(D \leqslant T_c) \leqslant \lambda\tau^2\|f\|_\infty.$$

On the other hand,

$$\int_0^{D\wedge T_c} g(Y_t)dt = \int_0^{D\wedge T_c'} g(Y_t')dt$$

so

$$\left|\mathbb{E}\int_0^{T_c} g(Y_t)dt - \mathbb{E}\int_0^{T_c'} g(Y_t')dt\right| \leqslant 2\lambda\tau^2\|f\|_\infty = \nu\tau^2\|f\|_\infty\|\mu_x - \mu_{x'}\|.$$

We apply this estimate with $g = f(x, \cdot)$ to obtain

$$\begin{aligned}|\chi(x,y)-\chi(x',y)| &= \left|\mathbb{E}\int_0^{T_c}(f(x,\bar{Y}_t)-f(x,Y_t))dt - \mathbb{E}\int_0^{T_c'}(f(x',\bar{Y}_t')-f(x',Y_t'))dt\right| \\ &\leqslant 2\tau\sup_{z\in I}|f(x,z)-f(x',z)| + \left|\mathbb{E}\int_0^{T_c} f(x,\bar{Y}_t)dt - \mathbb{E}\int_0^{T_c'} f(x,\bar{Y}_t')dt\right| \\ &\qquad + \left|\mathbb{E}\int_0^{T_c} f(x,Y_t)dt - \mathbb{E}\int_0^{T_c'} f(x,Y_t')dt\right| \\ &\leqslant 2\tau\sup_{z\in I}|f(x,z)-f(x',z)| + 2\nu\tau^2\|f\|_\infty\|\mu_x-\mu_{x'}\|\end{aligned}$$

as required. $\square$

To summarize, we have shown the following:

Proposition 1.4. *Assume conditions (9) and (10). Then, for any bounded measurable function f on $U \times I$, there exists a corrector χ_f for f satisfying the estimates (12) and (13).*

1.3 Statement of the estimates

Recall the context of Subsection 1.1. We consider a continuous-time Markov chain X with countable state-space S and generator matrix Q. We choose fluid coordinates $\boldsymbol{x} : S \to \mathbb{R}^d$ and an auxiliary coordinate $y : S \to I$. We choose also a subset $U \subseteq \mathbb{R}^d$, which provides a means of localization, together with a map $b : U \times I \to \mathbb{R}^d$, and a family $G = (G_x : x \in U)$ of generator matrices on I, each having a unique invariant distribution π_x. Choose also, as in the preceding subsection, a coupling mechanism for G. This comprises a constant $\nu > 0$, an auxiliary space E, a function $F : I \times E \to I$ and a family of probability distributions $\mu = (\mu_x : x \in U)$ on E such that

$$g(x,y,y') = \nu\mu_x(\{v \in E : F(y,v) = y'\}), \quad x \in U, \quad y, y' \in I \text{ distinct.}$$

Define for $x \in U$

$$\bar{b}(x) = \sum_{y\in I} b(x,y)\pi(x,y).$$

Write $\boldsymbol{X}_t = \boldsymbol{x}(X_t)$ and assume that $(x_t)_{0\leqslant t\leqslant t_0}$ is a solution in U to $\dot{x}_t = \bar{b}(x_t)$. We use a scaled supremum norm on $\mathbb{R}^d$: fix positive constants $\sigma_1, \ldots, \sigma_d$ and define for $x \in \mathbb{R}^d$

$$\|x\| = \max_{1\leqslant i\leqslant d}|x_i|/\sigma_i.$$

We introduce now some constants $\Lambda, B, \tau, J, J_1(b), J(\mu), K$ which characterize certain regularity properties of Q, b, and G. Assume that, for all $\xi \in S$, all $x \in U$ and all $y, y' \in I$,

$$q(\xi) \leqslant \Lambda, \quad \|b(x,y)\| \leqslant B, \quad m(x,y,y') \leqslant \tau. \tag{14}$$

Here $m(x,y,y')$ is the mean coupling time for G_x starting from y and y', defined in the preceding subsection, which depends on the choice of coupling mechanism. Write $\mathcal{J}$ for the set of pairs of points in U between which $\boldsymbol{X}$ can jump, thus

$$\mathcal{J} = \{(x,x') \in U \times U : x = \boldsymbol{x}(\xi), x' = \boldsymbol{x}(\xi') \text{ for some } \xi, \xi' \in S \text{ with } q(\xi,\xi') > 0\}.$$

Set

$$J = \sup_{(x,x') \in \mathcal{J}} \|x - x'\|, \quad J_1(b) = \sup_{(x,x') \in \mathcal{J},\, y \in I} \|b(x,y) - b(x',y)\|, \quad J(\mu) = \sup_{(x,x') \in \mathcal{J}} \|\mu_x - \mu_{x'}\|.$$

Write K for the Lipschitz constant of $\bar{b}$ on U: thus, for all $x, x' \in U$,

$$\|\bar{b}(x) - \bar{b}(x')\| \leqslant K\|x - x'\|. \tag{15}$$

Recall from Subsection 1.1 the definitions of the drift vector β for $\boldsymbol{x}$ and the jump rate γ for y. Define

$$T = \inf\{t \geqslant 0 : \boldsymbol{X}_t \notin U\}.$$

Fix constants $\delta(\beta,b), \delta(\gamma,g) \in (0,\infty)$ and consider the events

$$\Omega(\beta,b) = \left\{ \int_0^{T \wedge t_0} \|\beta(X_t) - b(\boldsymbol{x}(X_t), y(X_t))\| dt \leqslant \delta(\beta,b) \right\} \tag{16}$$

and

$$\Omega(\gamma,g) = \left\{ \int_0^{T \wedge t_0} \sum_{y' \neq y(X_t)} |\gamma(X_t,y') - g(\boldsymbol{x}(X_t), y(X_t), y')| dt \leqslant \delta(\gamma,g) \right\}. \tag{17}$$

Theorem 1.5. *Let $\varepsilon > 0$ be given and set $\delta = \varepsilon e^{-Kt_0}/7$. Assume that $J \leqslant \varepsilon$ and*

$$\max\left\{\|\boldsymbol{X}_0 - x_0\|,\, \delta(\beta,b),\, 2\tau B\delta(\gamma,g),\, 2\tau B,\, 2\Lambda t_0(\tau J_1(b) + \nu\tau^2 BJ(\mu))\right\} \leqslant \delta.$$

Set $\bar{J} = J + 4\tau B$ and assume that $\delta \leqslant \Lambda \bar{J} t_0/4$. Assume further that the following tube condition *holds:*

$$\text{for } \xi \in S \text{ and } t \leqslant t_0,\ \|\boldsymbol{x}(\xi) - x_t\| \leqslant 2\varepsilon \implies \boldsymbol{x}(\xi) \in U.$$

Then

$$\mathbb{P}\left(\sup_{t \leqslant t_0} \|\boldsymbol{X}_t - x_t\| > \varepsilon\right) \leqslant 2de^{-\delta^2/(4\Lambda\bar{J}^2 t_0)} + \mathbb{P}(\Omega(\beta,b)^c \cup \Omega(\gamma,g)^c).$$

The reader will understand precisely the role of the inequalities which appear as hypotheses in this result by following the proof. Here is an informal guide to their meanings. The tube condition, together with $J \leqslant \varepsilon$, allows us to localize the other hypotheses to U by trapping the process inside a tube around the limit path: these conditions can be satisfied by choosing U sufficiently large. The conditions $\|\boldsymbol{X}_0 - x_0\| \leqslant \delta$ and $\delta(\beta, b) \leqslant \delta$ enforce that the initial conditions and drift fields match closely. This requires in particular that the fluid and auxiliary coordinates provide sufficient information to nearly determine β. The condition on $\delta(\gamma, g)$ forces a close match between the local behaviour of the fast variable and the idealized fast process used to compute the corrector. The condition $2\tau B \leqslant \delta$ allows us to control the size of the corrector, balancing the mean recurrence time of the fast variable τ against the range of the drift field b. The condition on $2\Lambda t_0(\tau J_1(b) + \nu\tau^2 BJ(\mu))$ is needed for local regularity of the corrector, allowing us to pass back from the idealized fast process at one point x to the actual fast variable when the fluid variable is near x. Finally the condition $\delta \leqslant \Lambda \bar{J} t_0/4$ ensures we are in the 'Gaussian regime' of the exponential martingale inequality, where bad events cannot occur by a small number of large jumps. For a non-trivial limiting dynamics, ΛJ should be of order 1, while for a useful estimate ΛJ^2 should be small – thus, as expected, we can attempt to use the result when the Markov chain takes small jumps at a high rate.

It is sometimes possible to improve on the constant $\Lambda \bar{J}^2$ appearing in the preceding estimate, thereby obtaining useful probability bounds for smaller choices of ε. However, to do this we have to make hypotheses expressed in terms of a corrector. Fix $\bar{y} \in I$ and denote by χ the corrector for b given by (11). Define for $\xi \in S$ with $\boldsymbol{x}(\xi) \in U$

$$\bar{\boldsymbol{x}}(\xi) = \boldsymbol{x}(\xi) - \chi(\boldsymbol{x}(\xi), y(\xi)).$$

Define, for $\xi \in S$ such that $\boldsymbol{x}(\xi) \in U$ and $\boldsymbol{x}(\xi') \in U$ whenever $q(\xi, \xi') > 0$,

$$\alpha^i(\xi) = \sum_{\xi' \neq \xi} \{\bar{\boldsymbol{x}}^i(\xi') - \bar{\boldsymbol{x}}^i(\xi)\}^2 q(\xi, \xi'), \quad i = 1, \ldots, d.$$

Note that, since we shall be interested only in upper bounds, we deal here only with the diagonal terms of the diffusivity tensor defined at (6). Choose functions $a^i : I \to [0, \infty)$ such that, for all $\xi \in U$ where $\alpha^i(\xi)$ is defined,

$$\alpha^i(\xi) \leqslant a^i(y(\xi)), \quad i = 1, \ldots, d. \tag{18}$$

For simplicity, we do not allow a to depend on the fluid variable $\boldsymbol{x}(\xi)$. Since we can localize our hypotheses near the (compact) limit path, we do not expect to lose much precision by this simpification. On the other hand, by permitting a dependence on the fast variable we can sometimes do significantly better than Theorem 1.5, as we shall see in Section 2. Set

$$\bar{a}(x) = \sum_{y \in I} a(y) \pi(x, y), \quad x \in U.$$

We introduce two further constants A and $\bar{A}$, with $\bar{A} \leqslant A \leqslant \Lambda \bar{J}^2$. Assume that, for all $x \in U$ and all $y \in I$,

$$a^i(y) \leqslant A\sigma_i^2, \quad \bar{a}^i(x) \leqslant \bar{A}\sigma_i^2, \quad i = 1, \ldots, d. \tag{19}$$

Note that the corrector bound (12) gives $\|\chi(\boldsymbol{x}(\xi), y(\xi))\| \leqslant 2\tau B$, so $\alpha^i(\xi) \leqslant \Lambda \bar{J}^2 \sigma_i^2$ and so (19) holds with $A = \bar{A} = \Lambda \bar{J}^2$ and $a^i(y) = A\sigma_i^2$ and $\bar{a}^i(x) = A\sigma_i^2$. Thus Theorem 1.5 follows directly from (20) below. The new inequalities required on the left side of (21) can be understood roughly as imposing that the ratio of the averaged diffusivity to a uniform bound on the diffusivity is not too small compared to the mean recurrence time of the fast variable – so an effective averaging takes place.

Theorem 1.6. *Assume that the hypotheses of Theorem 1.5 hold and that $\delta \bar{J} \leqslant At_0/4$. Then*

$$\mathbb{P}\left(\sup_{t \leqslant t_0} \|\boldsymbol{X}_t - x_t\| > \varepsilon\right) \leqslant 2de^{-\delta^2/(4At_0)} + \mathbb{P}(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c). \tag{20}$$

Moreover, under the further conditions $\delta \bar{J} \leqslant \bar{A}t_0/4$ and

$$\frac{1}{t_0} \max\left\{\tau,\, \tau\delta(\gamma, g),\, \Lambda t_0 \nu \tau^2 J(\mu)\right\} \leqslant \bar{A}/(20A) \leqslant \Lambda\tau \tag{21}$$

we have

$$\mathbb{P}\left(\sup_{t \leqslant t_0} \|\boldsymbol{X}_t - x_t\| > \varepsilon\right) \leqslant 2de^{-\delta^2/(4\bar{A}t_0)} + 2de^{-(\bar{A}/A)^2 t_0/(6400\Lambda\tau^2)} + \mathbb{P}(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c). \tag{22}$$

Proof. Consider the stopping time

$$T_0 = \inf\{t \geqslant 0 : \|\boldsymbol{X}_t - x_t\| > \varepsilon\}.$$

By the tube condition, we have $T_0 \leqslant T$. Moreover, for any $t < T_0$ and any $\xi' \in S$ such that $q(X_t, \xi') > 0$, we have

$$\|\boldsymbol{x}(\xi') - x_t\| \leqslant J + \|\boldsymbol{X}_t - x_t\| \leqslant 2\varepsilon$$

so by the tube condition $\boldsymbol{x}(\xi') \in U$.

Recall that χ is the corrector for b given by (11). For the proof of (22), we shall use (11) to construct also a corrector $\tilde{\chi}$ for a. Set $\tilde{\delta} = \bar{A}t_0/10$. Note from (12) the bounds

$$\|\chi(x, y)\| \leqslant 2\tau B \leqslant \delta, \quad |\tilde{\chi}^i(x, y)| \leqslant 2\tau A\sigma_i^2 \leqslant \tilde{\delta}\sigma_i^2.$$

The inequality involving $\tilde{\delta}$ and further such inequalities below, which depend on the first inequality in assumption (21), will not be used in the proof of (20). Write $\Delta = G\chi(\boldsymbol{x}, y) - Q(\chi(\boldsymbol{x}, y)) = \Delta_1 + \Delta_2$ and $\tilde{\Delta} = G\tilde{\chi}(\boldsymbol{x}, y) - Q(\tilde{\chi}(\boldsymbol{x}, y)) = \tilde{\Delta}_1 + \tilde{\Delta}_2$, where

$$\Delta_1(\xi) = \sum_{y' \neq y(\xi)} \{g(\boldsymbol{x}(\xi), y(\xi), y') - \gamma(\xi, y')\}\chi(\boldsymbol{x}(\xi), y') \tag{23}$$

and

$$\Delta_2(\xi) = \sum_{\xi' \neq \xi} q(\xi, \xi')\{\chi(\boldsymbol{x}(\xi), y(\xi')) - \chi(\boldsymbol{x}(\xi'), y(\xi'))\} \tag{24}$$

and where $\tilde{\Delta}_1$ and $\tilde{\Delta}_2$ are defined analogously. Then, on $\Omega(\gamma, g)$, for $t \leqslant T \wedge t_0$,

$$\left\| \int_0^t \Delta_1(X_s) ds \right\| \leqslant 2\tau B \delta(\gamma, g) \leqslant \delta$$

and, using Proposition 1.3,

$$\left\| \int_0^t \Delta_2(X_s) ds \right\| \leqslant 2\Lambda t_0(\tau J_1(b) + \nu\tau^2 B J(\mu)) \leqslant \delta.$$

Similarly, for $t \leqslant T \wedge t_0$,

$$\left| \int_0^t \tilde{\Delta}_1^i(X_s) ds \right| \leqslant 2\tau A \delta(\gamma, g)\sigma_i^2 \leqslant \tilde{\delta}\sigma_i^2$$

and

$$\left| \int_0^t \tilde{\Delta}_2^i(X_s) ds \right| \leqslant 2\Lambda t_0 \nu \tau^2 A J(\mu) \sigma_i^2 \leqslant \tilde{\delta}\sigma_i^2.$$

Take $M = M^{\bar{\boldsymbol{x}}}$ as in equations (3) and (5) and consider the event

$$\Omega(M) = \left\{ \sup_{t \leqslant T_0 \wedge t_0} \|M_t\| \leqslant \delta \right\}.$$

Then, on $\Omega(\beta, b) \cap \Omega(\gamma, g) \cap \Omega(M)$, we can estimate the terms in equation (5) to obtain for $t \leqslant T_0 \wedge t_0$,

$$\|\boldsymbol{X}_t - x_t\| \leqslant 7\delta + K \int_0^t \|\boldsymbol{X}_s - x_s\| ds$$

so that $\|\boldsymbol{X}_t - x_t\| \leqslant \varepsilon$ by Gronwall's lemma. Note that this forces $T_0 \geqslant t_0$ and hence $\sup_{t \leqslant t_0} \|\boldsymbol{X}_t - x_t\| \leqslant \varepsilon$. Set $\rho = 3\bar{A}/2$ and consider the event

$$\Omega(a) = \left\{ \int_0^{T_0 \wedge t_0} a^i(Y_s) ds \leqslant \rho t_0 \sigma_i^2 \quad \text{for} \quad i = 1, \dots, d \right\}.$$

By condition (18), on $\Omega(a)$ we have

$$\int_0^{T_0 \wedge t_0} \alpha^i(X_s) ds \leqslant \rho t_0 \sigma_i^2.$$

Set

$$J_i = J(\bar{\boldsymbol{x}}^i) = \sup_{\xi, \xi' \in S,\, \boldsymbol{x}(\xi), \boldsymbol{x}(\xi') \in U,\, q(\xi, \xi') > 0} |\bar{\boldsymbol{x}}^i(\xi) - \bar{\boldsymbol{x}}^i(\xi')|, \quad i = 1, \dots, d$$

and use (12) to see that $J_i \leqslant \bar{J}\sigma_i$. Determine $\theta_i \in (0,\infty)$ by $\theta_i e^{\theta_i J_i} = \delta/(\rho t_0 \sigma_i)$; then $\theta_i \leqslant \delta/(\rho t_0\sigma_i)$, so $\theta_i J_i \leqslant 2\delta\bar{J}/(3\bar{A}t_0) \leqslant 1/4$, since we assumed that $\delta\bar{J} \leqslant \bar{A}t_0/4$. Since $e^{1/4} \leqslant 4/3$, we have $\rho e^{\theta_i J_i} \leqslant 2\bar{A}$. We now apply the exponential martingale inequality, Proposition 1.1, substituting $\pm\bar{\boldsymbol{x}}^i$ for ϕ for $i = 1,\dots,d$ and substituting $\delta\sigma_i$ for δ and $\rho t_0\sigma_i^2$ for ε. We thus obtain

$$\mathbb{P}(\Omega(M)^c \cap \Omega(a)) \leqslant 2de^{-\delta^2/(4\bar{A}t_0)}.$$

If we take $\bar{A} = A$, then, using (18) and (19), we have $\Omega(a) = \Omega$, so the proof of (20) is now complete.

Set $\eta = 16\Lambda\tau^2A^2$. We shall complete the proof of (22) by showing that

$$\mathbb{P}(\Omega(a)^c \cap \Omega(\gamma, g)) \leqslant 2de^{-\tilde{\delta}^2/(4\eta t_0)}.$$

Take $\tilde{M}$ as in equation (8), with a as in (18). Then, for $t \leqslant T$,

$$\int_0^t a(Y_s)ds = \tilde{\chi}(\boldsymbol{X}_t, Y_t) - \tilde{\chi}(\boldsymbol{X}_0, Y_0) - \tilde{M}_t + \int_0^t \tilde{\Delta}(X_s)ds + \int_0^t \bar{a}(\boldsymbol{X}_s)ds \tag{25}$$

where $\tilde{\Delta} = G\tilde{\chi}(\boldsymbol{x}, y) - Q(\tilde{\chi}(\boldsymbol{x}, y))$. Consider the event

$$\Omega(\tilde{M}) = \left\{ \sup_{t\leqslant T_0\wedge t_0} |\tilde{M}_t^i| \leqslant \tilde{\delta}\sigma_i^2 \quad \text{for} \quad i = 1,\dots,d \right\}.$$

Then, on $\Omega(\gamma, g) \cap \Omega(\tilde{M})$, we can estimate the terms in equation (25) to obtain

$$\int_0^{T_0\wedge t_0} a^i(Y_s)ds \leqslant (5\tilde{\delta} + \bar{A}t_0)\sigma_i^2 \leqslant \rho t_0\sigma_i^2.$$

Hence it will suffice to show that

$$\mathbb{P}(\Omega(\tilde{M})^c) \leqslant 2de^{-\tilde{\delta}^2/(4\eta t_0)}.$$

For this, we use again the exponential martingale inequality. Take $\phi(\xi) = \pm\tilde{\chi}^i(\boldsymbol{x}(\xi), y(\xi))$ in Proposition 1.1 and note that $\alpha^\phi(\xi) \leqslant 16\Lambda\tau^2A^2\sigma_i^4$, so

$$\int_0^{T_0\wedge t_0} \alpha^\phi(X_s)ds \leqslant 16\Lambda\tau^2A^2\sigma_i^4 t_0 = \eta t_0\sigma_i^4.$$

Set

$$\tilde{J}_i = J(\phi) = \sup_{\xi,\xi'\in S,\, \boldsymbol{x}(\xi),\boldsymbol{x}(\xi')\in U,\, q(\xi,\xi')>0} |\phi(\xi) - \phi(\xi')|, \quad i = 1,\dots,d$$

then $\tilde{J}_i \leqslant 4\tau A\sigma_i^2$. Determine $\tilde{\theta}_i \in (0,\infty)$ by $\tilde{\theta}_i e^{\tilde{\theta}_i\tilde{J}_i} = \tilde{\delta}/(\eta t_0\sigma_i^2)$. Then $\tilde{\theta}_i \leqslant \tilde{\delta}/(\eta t_0\sigma_i^2)$ so $\tilde{\theta}_i\tilde{J}_i \leqslant \bar{A}/(40\Lambda\tau A) \leqslant 1/2$ and so $e^{\tilde{\theta}_i\tilde{J}_i} \leqslant 2$. Hence

$$\mathbb{P}(\Omega(\tilde{M})^c) \leqslant 2d\exp\left\{-\tilde{\delta}^2/(2\eta t_0 e^{\tilde{\theta}_i\tilde{J}_i})\right\} \leqslant 2de^{-\tilde{\delta}^2/(4\eta t_0)}$$

as required. $\square$

2 The supermarket model with memory

The supermarket model with memory is a variant, introduced in [8], of the 'join the shortest queue' model, which has been widely studied [9, 3, 4, 5, 6, 7]. We shall verify rigorously the asymptotic picture for large numbers of queues derived in [8]. This will serve as an example to illustrate the general theory of the preceding sections. The explicit form of the error probabilities in Theorem 1.6 is used to advantage in dealing with the infinite-dimensional character of the limit model.

Fix $\lambda \in (0,1)$ and an integer $n \geqslant 1$. We shall consider the limiting behaviour as $N \to \infty$ of the following queueing system. Customers arrive as a Poisson process of rate $N\lambda$ at a system of N single-server queues. At any given time, the length of one of the queues is kept under observation. This queue is called the *memory queue*. On each arrival, an independent random sample of size n is chosen from the set of all N queues. For simplicity, we sample with replacement, allowing repeats and allowing the choice of the memory queue. The customer joins whichever of the memory queue or the sampled queues is shortest, choosing randomly in the event of a tie. Immediately after the customer has joined a queue, we switch the memory queue, if necessary, so that it is the currently shortest queue among the queues just sampled and the previous memory queue. The service requirements of all customers are assumed independent and exponentially distributed of mean 1.

Write $Z_t^k = Z_t^{N,k}$ for the proportion of queues having at least k customers at time t, and write Y_t for length of the memory queue at time t. Set $Z_t = (Z_t^k : k \in \mathbb{N})$ and $X_t = (Z_t, Y_t)$. Then $X = (X_t)_{t \geqslant 0}$ is a Markov chain, taking values in $S = S_0 \times \mathbb{Z}^+$, where S_0 is the set of non-increasing sequences in $N^{-1}\{0, 1, \ldots, N\}$ with finitely many non-zero terms. We shall treat Y as a fast variable and prove a fluid limit for Z as $N \to \infty$.

2.1 Statement of results

Let D be the set of non-increasing sequences[6] $z = (z_k : k \in \mathbb{N})$ in the interval $[0,1]$ such that

$$m(z) := \sum_k z_k < \infty.$$

Define for $z \in D$ and $k \in \mathbb{N}$

$$\mu(z,k) = \prod_{j=1}^{k} \frac{z_j^n}{1 - p_{j-1}(z)} \tag{26}$$

where

$$p_{k-1}(z) = n(z_{k-1} - z_k) z_k^{n-1}$$

[6]To lighten the notation, we shall sometimes move the coordinate index from a superscript to a subscript, allowing the nth power of the kth coordinate to be written z_k^n. We shall also write the time variable sometimes as a subscript, sometimes as an argument.

and where we take $z_0 = 1$. Set $\mu(z, 0) = 1$ for all z. An elementary calculation shows that in the case $n \geqslant 2$

$$p_{k-1}(z) \leqslant (1 - 1/n)^{n-1} \leqslant e^{-1/2} < 1. \tag{27}$$

In the case $n = 1$ we have $p_{k-1}(z) = z_{k-1} - z_k \leqslant 1$ and it is possible that $0/0$ appears in the product (26). For definiteness we agree to set $0/0 = 1$ in this case. Note that $\mu(z, k) \geqslant \mu(z, k+1)$ for all $k \geqslant 0$. Define for $z \in D$

$$v_k(z) = \lambda z_{k-1}^n \mu(z, k-1) - \lambda z_k^n \mu(z, k) - (z_k - z_{k+1})$$

and consider the differential equation

$$\dot{z}(t) = v(z(t)), \quad t \geqslant 0. \tag{28}$$

By a *solution* to (28) in D we mean a family of differentiable functions $z_k : [0, \infty) \to [0, 1]$ such that for all $t \geqslant 0$ and $k \in \mathbb{N}$ we have $(z_k(t) : k \in \mathbb{N}) \in D$ and

$$\dot{z}_k(t) = v_k(z(t)).$$

Theorem 2.1. *For all $z(0) \in D$, the differential equation $\dot{z}(t) = v(z(t))$ has a unique solution in D starting from $z(0)$. Moreover, if $(w(t) : t \in D)$ is another solution in D with $z_k(0) \leqslant w_k(0)$ for all k, then $z_k(t) \leqslant w_k(t)$ for all k and all $t \geqslant 0$.*

There is a fixed point of these dynamics $a \in D$ given by setting $a_0 = 1$ and defining

$$a_{k+1} = \lambda a_k^n \mu(a, k), \quad k \geqslant 0. \tag{29}$$

The components of a decay super-geometrically. Set

$$\alpha = n + \frac{1}{2} + \sqrt{n^2 + \frac{1}{4}}. \tag{30}$$

Then $\alpha \in (2n, 2n+1)$.

Theorem 2.2. *We have*

$$\lim_{k \to \infty} \frac{1}{k} \log\log\left(\frac{1}{a_k}\right) = \alpha.$$

Assume for simplicity that we start the queueing system from the state where all queues except the memory queue are empty and where the memory queue has exactly one customer. Write $(z(t) : t \geqslant 0)$ for the solution to (28) starting from 0. Then $z_k(t) \leqslant a_k$ for all k and t. Our main result shows that $(z(t) : t \geqslant 0)$ is a good approximation to the process of empirical distributions of queue lengths $(Z^N(t) : t \geqslant 0)$ for large N. The sense of this approximation is reasonably sharp. In particular, as a straightforward corollary, we obtain that, on a given time interval $[0, t_0]$, for any $r > \alpha^{-1}$, with high probability as $N \to \infty$, no queue length exceeds $r \log\log N$.

Theorem 2.3. *Set $\kappa = (2\alpha)^{-1}$ and define*

$$d = d(N) = \sup\{k \in \mathbb{N} : Na_k > N^\kappa\}.$$

Fix a function ϕ on $\mathbb{N}$ such that $\phi(N)/N^\kappa \to 0$ and $\log \phi(N)/\log\log N \to \infty$ as $N \to \infty$. Set $\rho = 4/(1-\lambda)$ when $n = 1$ and set $\rho = 2^n/(1-e^{-1/2})$ when $n \geqslant 2$. Set $\tilde{a}_{d+1} = N^{-1}a_d^n + \rho^d a_{d+1}$. Then

$$\lim_{N\to\infty} d(N)/\log\log N = 1/\alpha. \tag{31}$$

Moreover, for all $t_0 \geqslant 0$, we have

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\leqslant t_0}\sup_{k\leqslant d}\frac{|Z_t^{N,k} - z_t^k|}{\sqrt{a_k}} \geqslant \sqrt{\frac{\phi(N)}{N}}\right) = 0 \tag{32}$$

and

$$\lim_{R\to\infty}\limsup_{N\to\infty} \mathbb{P}(Z_t^{N,d+1} \geqslant R\tilde{a}_{d+1} \text{ for some } t \leqslant t_0) = 0 \tag{33}$$

and

$$\lim_{N\to\infty} \mathbb{P}(Z_t^{N,d+2} = 0 \text{ for all } t \leqslant t_0) = 1. \tag{34}$$

The argument used to prove this result would apply without modification starting from any initial condition $z(0)$ for the limit dynamics (28) such that $z_k(0) \leqslant a_k$ for all k, with suitable conditions on the convergence of $Z^N(0)$ to $z(0)$. It may be harder to move beyond initial conditions which do not lie below the fixed point. We do note here however a family of long-time upper bounds for the limit dynamics which might prove useful for such an extension. Fix $j \in \mathbb{N}$ and define $a_k^{(j)} = a_{(k-j)^+}$ for each $k \in \mathbb{Z}^+$; then $a^{(j)}$ is a fixed point of the modified equation

$$\dot{w}_k(t) = v_k(w(t)) + (w_j(t) - w_{j+1}(t))1_{\{k=j\}}.$$

Since the added term is always non-negative, a similar argument to that used to prove Theorem 2.1 in the next subsection shows also that, if $z(0) \leqslant a^{(j)}$ and $(z(t) : t \geqslant 0)$ is a solution of the original equation, then $z(t) \leqslant a^{(j)}$ for all t.

2.2 Existence and monotonicity of the limit dynamics

The differential equation (28) characterizes the limit dynamics for the fluid variables in our queueing model. Our analysis of its space of solutions will rest on the exploitation of certain non-negativity properties which have a natural probabilistic interpretation.

We consider first a truncated, finite-dimensional system. Fix $d \in \mathbb{N}$ and define a vector field $u = u^{(d)}$ on D by setting $u_k(z) = v_k(z)$ for $k \leqslant d-1$ and

$$u_d(z) = \lambda z_{d-1}^n \mu(z, d-1) - \lambda z_d^n \mu(z,d) - z_d \tag{35}$$

and $u_k(z) = 0$ for $k \geqslant d+1$. Set $D(d) = \{(x_1, \ldots, x_d, 0, 0\ldots) : 0 \leqslant x_d \leqslant \ldots \leqslant x_1 \leqslant 1\}$.

Proposition 2.4. *For all $x(0) \in D(d)$, the differential equation $\dot{x}(t) = u(x(t))$ has a unique solution $(x(t) : t \geqslant 0)$ in $D(d)$ starting from $x(0)$.*

Proof. In the proof, we consider $D(d)$ as a subset of $\mathbb{R}^d$. The function u is continuous on $D(d)$ and is differentiable in the interior of $D(d)$ with bounded partial derivatives. (In the case $n = 1$, the singularity in $(\partial/\partial x_j)\mu(x, k)$ for $j \leqslant k$ as $x_{j-1} - x_j \to 1$ is cancelled by the factor x_k by which it is multiplied, since $x_k \leqslant x_j$ on $D(d)$.) For $x \in D(d)$ we have $u_1(x) \leqslant 0$ when $x_1 = 1$, and $u_d(x) \geqslant 0$ when $x_d = 0$. Moreover, for $k = 1, \dots, d-1$, if $x_k = x_{k+1}$ then $p_k(x) = 0$ so

$$u_{k+1}(x) = x_k^n(\mu(x,k)-\mu(x,k+1)) \leqslant \mu(x,k)-\mu(x,k+1) \leqslant x_{k-1}^n\mu(x,k-1)-x_k^n\mu(x,k) \leqslant u_k(x).$$

The conclusion now follows by standard arguments. □

Proof of Theorem 2.1. Suppose that $(w(t) : t \geqslant 0)$ is a solution to $\dot{w}(t) = v(w(t))$ in D starting from $w(0)$, with $z(0) \leqslant w(0)$, that is to say $z_k(0) \leqslant w_k(0)$ for all k. Fix d and write $x(t) = z^{(d)}(t)$ for the solution to $\dot{x}(t) = u^{(d)}(x(t))$ in $D(d)$ starting from $(z_1(0), \dots, z_d(0), 0, 0, \dots)$. Set $y(t) = (w_1(t), \dots, w_d(t), 0, 0, \dots)$ and note that $x(0) \leqslant y(0)$ and $y(t) \in D(d)$ for all t. We shall show that $x(t) \leqslant y(t)$ for all t. Consider now $D(d)$ as a subset of $\mathbb{R}^d$. We have

$$\dot{y}(t) = u(y(t)) + w^{d+1}(t)e_d$$

where $e_d = (0, \dots, 0, 1)$. Note that $w^{d+1}(t) \geqslant 0$ for all t. Now u is Lipschitz on $D(d)$ and for $k = 1, \dots, d$ we can show that[7]

$$x, y \in D(d), \quad x \leqslant y, \quad x_k = y_k \quad \Longrightarrow \quad u_k(x) \leqslant u_k(y).$$

Hence by a standard argument $z^{(d)}(t) = x(t) \leqslant y(t) \leqslant w(t)$ for all t. The same argument shows that $z^{(d)}(t) \leqslant z^{(d+1)}(t)$ for all t, so the limit $z_k(t) = \lim_{d\to\infty} z_k^{(d)}(t)$ exists for all k and t, and $z(t) \leqslant w(t)$ for all t.

Fix k and take $d \geqslant k+1$. Then the following equation holds for all t

$$\begin{aligned} z_k^{(d)}(t) &+ \int_0^t \lambda z_k^{(d)}(s)^n \mu(z^{(d)}(s), k)ds + \int_0^t z_k^{(d)}(s)ds \\ &= z_k(0) + \int_0^t \lambda z_{k-1}^{(d)}(s)^n \mu(z^{(d)}(s), k-1)ds + \int_0^t z_{k+1}^{(d)}(s)ds. \end{aligned}$$

[7] An elementary calculation shows that $\mu(x, j) \leqslant \mu(y, j)$ for all j whenever $x \leqslant y$. This will also be shown by a soft probabilistic argument in Subsection 2.6. The further condition $x_k = y_k$ gives the inequality

$$y_{k-1}^n - x_{k-1}^n \geqslant n(y_{k-1} - x_{k-1})x_k^{n-1} = p_{k-1}(y) - p_{k-1}(x) \geqslant \frac{y_k^{2n}}{1 - p_{k-1}(y)} - \frac{x_k^{2n}}{1 - p_{k-1}(x)}.$$

On letting $d \to \infty$, we see by monotone convergence that

$$\begin{aligned} z_k(t) &+ \int_0^t \lambda z_k(s)^n \mu(z(s),k)ds + \int_0^t z_k(s)ds \\ &= z_k(0) + \int_0^t \lambda z_{k-1}(s)^n \mu(z(s),k-1)ds + \int_0^t z_{k+1}(s)ds. \end{aligned}$$

Since $z(t) \in D$ for all t, all integrands in this equation are bounded by 1. It is now straightforward to see that $(z(t) : t \geqslant 0)$ is a solution.

Now

$$\begin{aligned} w_k(t) &+ \int_0^t \lambda w_k(s)^n \mu(w(s),k)ds + \int_0^t w_k(s)ds \\ &= w_k(0) + \int_0^t \lambda w_{k-1}(s)^n \mu(w(s),k-1)ds + \int_0^t w_{k+1}(s)ds. \end{aligned}$$

By summing these equations over $k \in \{1,\dots,d\}$ we see that the map $t \mapsto \sum_{k=1}^d w_k(t) - \lambda t$ is non-increasing for all d. Hence $m(w(t)) \leqslant m(w(0)) + \lambda t < \infty$. The equations can then be summed over all k and rearranged to obtain

$$m(w(t)) = m(w(0)) + \lambda t - \int_0^t w_1(s)ds.$$

On the other hand

$$m(z^{(d)}(t)) = m(z^{(d)}(0)) + \lambda t - \int_0^t z_1^{(d)}(s)ds - \lambda \int_0^t z_d^{(d)}(s)^n \mu(z^{(d)}(s),d)ds$$

so

$$m(w(t)) - m(z^{(d)}(t)) \leqslant m(w(0)) - m(z^{(d)}(0)) + \lambda \int_0^t z_d(s)^n \mu(z(s),d)ds. \qquad (36)$$

If $w(0) = z(0)$ then the right hand side tends to 0 as $d \to \infty$ so we must have $z(t) = w(t)$ for all t. □

2.3 Properties of the fixed point

Recall the definition (29) of the fixed point a. Since $\mu(a,k) \leqslant 1$ for all k, we have

$$a_k \leqslant \lambda^{1+n+\dots+n^{k-1}}$$

so $a_k \to 0$ as $k \to \infty$. Theorem 2.2 is then a straightforward corollary of the following estimate.

Proposition 2.5. *There is a constant $C(\lambda, n) < \infty$ such that, for all $k \geqslant 0$,*

$$C^{-1} a_k^\alpha \leqslant a_{k+1} \leqslant C a_k^\alpha \tag{37}$$

where α is given by equation (30).

Proof. Note that, since $a_1 = \lambda$, we have $p_{k-1}(a) = a_{k-1} - a_k \leqslant \lambda \vee (1-\lambda) < 1$ for all k when $n = 1$. On the other hand equation (27) gives $p_{k-1}(z) \leqslant e^{-1/2}$ for all k when $n \geqslant 2$. Then from $\sum_{k=1}^\infty p_{k-1}(a) \leqslant 1$, we obtain a constant $c < \infty$, which may depend on λ when $n = 1$, such that

$$\prod_{k=1}^\infty \frac{1}{1 - p_{k-1}(a)} \leqslant c.$$

Then for $k \geqslant 0$

$$\lambda a_k^{2n} \prod_{j=1}^{k-1} a_j^n \leqslant a_{k+1} \leqslant c\lambda a_k^{2n} \prod_{j=1}^{k-1} a_j^n$$

so for $k \geqslant 1$

$$c^{-1} a_k^{2n+1} a_{k-1}^{-n} \leqslant a_{k+1} \leqslant c a_k^{2n+1} a_{k-1}^{-n}. \tag{38}$$

Note that $\lambda a_0^\alpha = \lambda = a_1 \leqslant \lambda^{-1} a_0^\alpha$. Fix $A \geqslant 1/\lambda$ and suppose inductively that

$$A^{-1} a_{k-1}^\alpha \leqslant a_k \leqslant A a_{k-1}^\alpha.$$

On using these inequalities to estimate a_{k-1} in (38), we obtain

$$(cA^{n/\alpha})^{-1} a_k^\alpha \leqslant a_{k+1} \leqslant cA^{n/\alpha} a_k^\alpha$$

where we have used the fact that $1 - n/\alpha = \alpha - 2n$. Hence the induction proceeds provided we take $A \geqslant c^{\alpha/(\alpha-n)}$. □

2.4 Choice of fluid coordinates and fast variable

In the remaining subsections we apply Theorem 1.6 to deduce Theorem 2.3. Define d as in Theorem 2.3 and take as auxiliary space $I = \mathbb{N}$ when $n = 1$ and $I = \mathbb{Z}^+$ when $n \geqslant 2$. Make the following choice of fluid and auxiliary coordinates: for $\xi = (z, y) \in S$ with $z = (z_k : k \in \mathbb{N})$, set

$$x^k(\xi) = z_k, \quad k = 1, \dots, d, \quad y(\xi) = y.$$

Thus our fluid variable is $\boldsymbol{X}_t = \boldsymbol{x}(X_t) = (Z_t^1, \dots, Z_t^d)$ and our fast variable is $Y_t = y(X_t)$. Note that when $n = 1$, if $Y_0 \geqslant 1$ then $Y_t \geqslant 1$ for all t, so Y takes values in $I = \mathbb{N}$.

Let us compute the drift vector $\beta(\xi)$ for $\boldsymbol{X}$ when X is in state $\xi = (z, y) \in S$. Note that X^k makes a jump of size $1/N$ when a customer arrives at a queue of length $k-1$, and makes a jump of size $-1/N$ when a customer departs from a queue of length k, otherwise

X^k is constant. The length of the queue which an arriving customer joins depends on the length of the memory queue y and on the lengths of the sampled queues. Denote the vector of sampled queue lengths by $V = V(z) = (V_1, \ldots, V_n)$ and write $V^{(1)} \leqslant V^{(2)} \leqslant \ldots \leqslant V^{(n)}$ for the ordered queue lengths. Define $\min(v) = v_1 \wedge \cdots \wedge v_n$ and set $M = \min(V)$. Then $M = V^{(1)}$ and

$$\mathbb{P}(M \geqslant k) = z_k^n.$$

A new customer will go to a queue of length at least k if and only if $M \geqslant k$ and $y \geqslant k$. So the rate for an arrival to a queue of length exactly $k-1$ is

$$N\lambda\mathbb{P}(M \geqslant k-1)1_{\{y \geqslant k-1\}} - N\lambda\mathbb{P}(M \geqslant k)1_{\{y \geqslant k\}}.$$

The rate for a departure from a queue of length k is $N(z_k - z_{k+1})$. Hence, setting $z_0 = 1$, we have

$$\beta_k(\xi) = \lambda z_{k-1}^n 1_{\{y \geqslant k-1\}} - \lambda z_k^n 1_{\{y \geqslant k\}} - (z_k - z_{k+1}).$$

We now compute (an approximation to) the jump rates $\gamma(\xi, y')$ for Y when X is in state $\xi = (z, y) \in S$. The rate of departures from the memory queue is at most 1. Arrivals to the system occur at rate $N\lambda$. Occasionally, the memory queue falls in the sample, an event of probability no greater than n/N and hence of rate no greater than λn. Assuming that the the memory queue does not fall in the sample, the length of the memory queue after an arrival is given by

$$F(y, V) = (y+1)1_{\{y \leqslant M-1\}} + y1_{\{M \leqslant y \leqslant P\}} + P1_{\{y \geqslant P+1\}} \tag{39}$$

where $P = p(V)$ is given by $P = M + 1$ when $n = 1$ and $P = (M+1) \wedge V^{(2)}$ otherwise. Hence we have

$$\sum_{y' \neq y(\xi)} |\gamma(\xi, y') - N\lambda\mathbb{P}(F(y, V(z)) = y')| \leqslant 1 + \lambda n. \tag{40}$$

2.5 Choice of limit characteristics and coupling mechanism

Define

$$U = \{x \in \mathbb{R}^d : 0 \leqslant x_d \leqslant \ldots \leqslant x_1 \leqslant 1 \text{ and } x_1 \leqslant (\lambda+1)/2 \text{ and } x_k \leqslant 2a_k \text{ for all } k\}.$$

The condition $x_1 \leqslant (\lambda+1)/2$ ensures that $1 - x_1$ is uniformly positive on U. Define $b : U \times \mathbb{Z}^+ \to \mathbb{R}^d$ by

$$b_k(x, y) = \lambda x_{k-1}^n 1_{\{y \geqslant k-1\}} - \lambda x_k^n 1_{\{y \geqslant k\}} - (x_k - x_{k+1}) \tag{41}$$

where we set $x_0 = 1$ and $x_{d+1} = 0$. Then, for $\xi \in S$ with $\boldsymbol{x}(\xi) \in U$ we have

$$\beta(\xi) = b(\boldsymbol{x}(\xi), y(\xi)) + (0, \ldots, 0, z_{d+1}). \tag{42}$$

It is convenient to specify our choice of the generator matrices $(G_x : x \in U)$ and our choice of coupling mechanism at the same time. Set $\nu = N\lambda$ and take as auxiliary space $E = (\mathbb{Z}^+)^n$. Define a family of probability distributions $\mu = (\mu_x : x \in U)$ on E, taking μ_x to be the law of a random sample $V = V(x) = (V_1, \dots, V_n)$ with

$$\mathbb{P}(V_1 \geqslant k) = \cdots = \mathbb{P}(V_n \geqslant k) = x_k$$

for $k = 0, 1, \dots, d+1$. Note that

$$\|\mu_x - \mu_{x'}\| \leqslant 2n \sum_{k=1}^{d} |x_k - x'_k|. \tag{43}$$

Then define for distinct $y, y' \in \mathbb{Z}^+$

$$g(x, y, y') = N\lambda \mathbb{P}(F(y, V(x)) = y')$$

where F is given by (39). We take as coupling mechanism the triple (ν, μ, F).

Note that $F(y, v) = F(\bar{y}, v)$ for all $y, \bar{y} \in I$ whenever $p(v) = \min I$. For $x \in U$ we have

$$\mathbb{P}\left(p(V(x)) = 1\right) \geqslant 1 - x_1 > \frac{1-\lambda}{2}$$

when $n = 1$, whereas for $n \geqslant 2$ we have

$$\mathbb{P}\left(p(V(x)) = 0\right) \geqslant (1 - x_1)^2 > \left(\frac{1-\lambda}{2}\right)^2$$

Hence we obtain, in all cases, $m(x, y, \bar{y}) \leqslant \tau$, where we set

$$\tau = \frac{4}{N\lambda(1-\lambda)^2}.$$

For $\xi \in S$ with $x = \boldsymbol{x}(\xi) \in U$ we can realize a sample $V(z)$ (from the distribution of queue lengths) and the sample $V(x)$ on the same probability space by setting $V_i(x) = V_i(z) \wedge d$. Write $M(x) = \min(V(x))$ and $P(x) = p(V(x))$. Then $M(x) = M(z) \wedge d$ and $P(x) = P(z) \wedge (d+1)$ when $n = 1$ and $P(x) = P(z) \wedge d$ when $n \geqslant 2$. The difference between the two cases is that there is no second shortest queue in the sample when $n = 1$. We have, for $n = 1$,

$$\mathbb{P}(P(z) \neq P(x)) \leqslant \mathbb{P}(M(z) \geqslant d+1) = z_{d+1}$$

and, for $n \geqslant 2$,

$$\mathbb{P}(P(z) \neq P(x)) \leqslant \mathbb{P}(P(z) \geqslant d+1) \leqslant n z_d z_{d+1}^{n-1}.$$

Now $P(x) = P(z)$ implies $M(x) = M(z)$ and hence $F(y, V(x)) = F(y, V(z))$ for all y. Hence

$$\mathbb{P}(F(y, V(z)) \neq F(y, V(x))) \leqslant \mathbb{P}(P(x) \neq P(z)).$$

On combining this with (40) we obtain

$$\sum_{y' \neq y(\xi)} |\gamma(\xi, y') - g(\boldsymbol{x}(\xi), y(\xi), y')| \leqslant 1 + \lambda n + N\lambda n z_{d+1}. \tag{44}$$

2.6 Local equilibrium distribution

The Markov chain determined by the generator G_x has a unique closed communicating class, which is contained in $\{0, 1, \ldots, d\}$. Hence G_x has a unique equilibrium distribution π_x, which is supported on $\{0, 1, \ldots, d\}$. Consider a continuous-time Markov chain[8] $Y = (Y_t)_{t \geqslant 0}$ with generator G_x, and initial distribution π_x. Set $\mu(x, k) = \mathbb{P}(Y_0 \geqslant k)$. Then Y jumps into $\{0, 1, \ldots, k\}$ from j at rate $\alpha = N\lambda(1 - x_{k+1}^n - p_k(x))$, for all $j \geqslant k + 1$. On the other hand, Y jumps out of $\{0, 1, \ldots, k\}$ only from k, and that at rate $\beta = N\lambda x_{k+1}^n$. Since the long run rates of such jumps must agree, we deduce that $\alpha\mu(x, k + 1) = \beta\pi_x(k)$. Hence we obtain

$$\mu(x, k + 1)(1 - p_k(x)) = x_{k+1}^n \mu(x, k)$$

and so

$$\mu(x, k) = \prod_{j=1}^{k} \frac{x_j^n}{1 - p_{j-1}(x)}, \quad k = 1, \ldots, d. \tag{45}$$

Hence our present notation is consistent with the definition (26). Note also that, for $z \in D$, $\mu(z, k)$ depends only on $z_1, \ldots, z_k$; in particular if $x = (z_1, \ldots, z_d)$ then $\mu(z, k) = \mu(x, k)$ for all $k \leqslant d$. Note that $\bar{b}$ is given by

$$\bar{b}_k(x) = \lambda x_{k-1}^n \mu(x, k - 1) - \lambda x_k^n \mu(x, k) - (x_k - x_{k+1}), \quad k = 1, \ldots, d \tag{46}$$

where $x_0 = 1$ and $x_{d+1} = 0$. Hence $\bar{b} = u^{(d)}$ as defined in Subsection 2.2.

A comparison of (35) and (41) now shows that $\bar{b} = u^{(d)}$.

Recall that $\rho = 4/(1 - \lambda)$ when $n = 1$ and that $\rho = 2^n/(1 - e^{-1/2})$ when $n \geqslant 2$. Then for $x \in U$ and $k \geqslant 1$ we have

$$\mu(x, k) = x_k^n \mu(x, k - 1)/(1 - p_{k-1}(x)) \leqslant \rho a_k^n \mu(x, k - 1) \leqslant \rho a_k^n \mu(x, k - 1)/(1 - p_{k-1}(a))$$

so, for all $x \in U$ and inductively for all $k \geqslant 1$, we obtain

$$\mu(x, k) \leqslant \rho^k \mu(a, k). \tag{47}$$

[8] See footnote 5.

The following argument shows that $\mu(z,k) \leqslant \mu(z',k)$ for all k whenever $z \leqslant z'$. Fix $d \geqslant k$ and set $x' = (z'_1, \dots, z'_d)$. Assume that $x \leqslant x'$. By a standard construction we can realize samples $V = V(x)$ and $V' = V(x')$ on a common probability space such that $V_i \leqslant V'_i$ for all i. Then we can construct Markov chains Y and Y', having generators G_x and $G_{x'}$ respectively, on the canonical space of a marked Poisson process of rate $N\lambda$, where the marks are independent copies of (V, V'), as follows. Set $Y_0 = Y'_0 = 1$ and define recursively at each jump time T of the Poisson process $Y_T = F(Y_{T-}, V_T)$ and $Y'_T = F(Y'_{T-}, V'_T)$, where (V_T, V'_T) is the mark at time T. Then, since F is non-decreasing in both arguments, we see by induction that $Y_t \leqslant Y'_t$ for all t. Hence, by convergence to equilibrium,

$$\mu(z,k) = \mu(x,k) = \lim_{t\to\infty} \mathbb{P}(Y_t \geqslant k) \leqslant \lim_{t\to\infty} \mathbb{P}(Y'_t \geqslant k) = \mu(x',k) = \mu(z',k).$$

2.7 Corrector upper bound

We take as our reference state $\bar{y} = \min I$ and note that, under the coupling mechanism, we have $\bar{Y}_t \leqslant Y_t$ for all t. Then the kth component of the corrector for b is given by

$$\chi_k(x,y) = \lambda \mathbb{E}_y \int_0^{T_c} (x_{k-1}^n 1_{\{\bar{Y}_s < k-1 \leqslant Y_s\}} - x_k^n 1_{\{\bar{Y}_s < k \leqslant Y_s\}}) ds$$

so for $x \in U$ and all $y \in I$ we have

$$|\chi_k(x,y)| \leqslant \tau x_{k-1}^n \leqslant C a_{k-1}^n / N.$$

Fix now $y \leqslant k-2$ and consider the stopping time $T = \inf\{t \geqslant 0 : Y_t = k-1\}$. Note that Y can enter state $k-1$ only from $k-2$ and does so at rate $N\lambda x_{k-1}^n$, whereas T_c occurs in state $k-2$ at rate at least $N\lambda(1-\lambda)^2/4$. Hence $\mathbb{P}_y(T \leqslant T_c) \leqslant C a_{k-1}^n$ and so, by the Markov property,

$$\mathbb{E}_y \int_0^{T_c} 1_{\{\bar{Y}_s < k-1 \leqslant Y_s\}} ds \leqslant \mathbb{E}_y(1_{\{T \leqslant T_c\}} m(x, k-1, \bar{Y}_T)) \leqslant C a_{k-1}^n \tau \leqslant C a_{k-1}^n / N.$$

Hence we obtain, for $x \in U$ and all $y \in I$,

$$|\chi_k(x,y)| \leqslant C(a_{k-1}^n 1_{\{y \geqslant k-1\}} + a_{k-1}^{2n})/N. \tag{48}$$

2.8 Quadratic variation upper bound

The growth rate at ξ of the quadratic variation of the corrected kth coordinate is given by

$$\alpha_k(\xi) = \sum_{\xi' \neq \xi} \{\bar{\boldsymbol{x}}_k(\xi') - \bar{\boldsymbol{x}}_k(\xi)\}^2 q(\xi, \xi').$$

Recall that $\bar{x}_k = x_k - \chi_k(\boldsymbol{x}, y)$. We estimate separately, writing $x = \boldsymbol{x}(\xi)$ and $y = y(\xi)$,

$$\sum_{\xi' \neq \xi} \{\boldsymbol{x}_k(\xi') - x_k\}^2 q(\xi, \xi') \leqslant N^{-2}(N x_k + N\lambda x_{k-1}^n 1_{\{y \geqslant k-1\}})$$

and

$$\sum_{\xi' \neq \xi} \{\chi_k(\boldsymbol{x}(\xi'), y(\xi')) - \chi_k(x, y)\}^2 q(\xi, \xi') \leqslant C N^{-1} x_{k-1}^{2n}.$$

When $y \leqslant k-2$ we can improve the last estimate by splitting the sum in two and using

$$\sum_{\xi' \neq \xi,\, y(\xi') \leqslant k-2} \{\chi_k(\boldsymbol{x}(\xi'), y(\xi')) - \chi_k(x, y)\}^2 q(\xi, \xi') \leqslant C N^{-1} x_{k-1}^{4n}$$

and

$$\sum_{\xi' \neq \xi,\, y(\xi') \geqslant k-1} \{\chi_k(\boldsymbol{x}(\xi'), y(\xi')) - \chi_k(x, y)\}^2 q(\xi, \xi') \leqslant C N^{-1} x_{k-1}^{3n}.$$

We used (48) for the first inequality and for the second used

$$\sum_{\xi' \neq \xi,\, y(\xi') \geqslant k-1} q(\xi, \xi') \leqslant C N x_{k-1}^n.$$

On combining these estimates we obtain

$$\alpha_k(\xi) \leqslant C(x_k + x_{k-1}^n 1_{\{y \geqslant k-1\}} + x_{k-1}^{3n}) \leqslant a_k(y(\xi))/N$$

where

$$a_k(y(\xi)) = C(a_k + a_{k-1}^n 1_{\{y \geqslant k-1\}})/N.$$

Then, using the estimate (47) and the limit (31), we have

$$\bar{a}_k(x) = C(a_k + a_{k-1}^n \mu(x, k-1)) \leqslant C\rho^{d-1} a_k/N \leqslant C(\log N)^C a_k/N$$

so we have for all $x \in U$ and $y \in I$

$$a_k(y) \leqslant A a_k, \quad \bar{a}_k(x) \leqslant \bar{A} a_k$$

with $A = C/(N a_d^{1-n/\alpha})$ and $\bar{A} = C(\log N)^C/N$. It is straightforward to check that, for N sufficiently large, we have $\bar{A} \leqslant A \leqslant \Lambda \bar{J}^2$.

2.9 Truncation estimates

A specific feature of the problem we consider is that the limit dynamics is infinite-dimensional, while the general fluid limit estimate applies in a finite-dimensional context. In this subsection we establish some truncation estimates which will allow us to reduce to finitely many dimensions.

Let $(z(t) : t \geqslant 0)$ be the solution in D to $\dot{z}(t) = v(z(t))$ starting from 0, as in Theorem 2.3. Let $(x(t) : t \geqslant 0)$ be the solution to $\dot{x}(t) = \bar{b}(x(t))$ starting from 0.

Lemma 2.6. *We have*

$$\sum_{k=1}^{d} |z_k(t) - x_k(t)| \leqslant t a_{d+1}.$$

Proof. Since $\bar{b} = u^{(d)}$, we have $x(t) = z^{(d)}(t)$ for all t, and so, from (36) we obtain

$$\sum_{k=1}^{d} |z_k(t) - x_k(t)| \leqslant \sum_{k=1}^{d} (z_k(t) - z_k^{(d)}(t)) \leqslant \lambda \int_0^t z_d(s)^n \mu(z(t), d) ds \leqslant t\lambda a_d^n \mu(a, d) = t a_{d+1}.$$

□

Denote by $A_k(t)$ the number of arrivals to queues of length at least k by time t. Note that $NZ_t^{k+1} \leqslant A_k(t)$ for all $k \geqslant 1$ and all t. Recall that

$$\tilde{a}_{d+1} = N^{-1} a_d^n + \rho^d a_{d+1}.$$

Lemma 2.7. *There is a constant $C(\lambda, n) < \infty$ such that, for all $t \geqslant 0$ and all N, we have*

$$\mathbb{E}(A_d(T \wedge t)) \leqslant C e^{Ct} N \tilde{a}_{d+1}.$$

Proof. Consider the function f on $U \times I$ given by $f(x, y) = 1_{\{y \geqslant d\}}$ and note that $\bar{f}(x) = \mu(x, d)$. Let χ be the corrector for f given by (11). Then, for all $x \in U$ and all $y \in I$

$$|\chi(x, y)| \leqslant 2\tau \|f\|_\infty = 2\tau = CN^{-1}$$

and, whenever $x = \boldsymbol{x}(\xi)$ and $x' = \boldsymbol{x}(\xi')$ with $q(\xi, \xi') > 0$, by the estimates (13) and (43),

$$|\chi(x, y) - \chi(x', y)| \leqslant 2\nu\tau^2 \|f\|_\infty \|\mu_x - \mu_{x'}\| \leqslant CN^{-2}. \tag{49}$$

By optional stopping,

$$\left| \int_0^{T \wedge t} Q(\chi(\boldsymbol{x}, y))(X_s) ds \right| = |\mathbb{E}(\chi(\boldsymbol{X}_{T\wedge t}, Y_{T\wedge t}) - \chi(\boldsymbol{X}_0, Y_0))| \leqslant CN^{-1}.$$

Now

$$Q(\chi(\boldsymbol{x}, y))(\xi) = 1_{\{y \geqslant d\}} - \mu(\boldsymbol{x}(\xi), d) - \Delta_1(\xi) - \Delta_2(\xi)$$

where Δ_1, Δ_2 are given by (23), (24). We use (44) to obtain the estimate

$$|\Delta_1(\xi)| \leqslant 2\tau(1 + \lambda n + N\lambda n z_{d+1}) = C(z_{d+1} + N^{-1})$$

and from (49) deduce that

$$|\Delta_2(\xi)| \leqslant N(1+\lambda)CN^{-2} = CN^{-1}.$$

So

$$\mathbb{E}\int_0^{T\wedge t} 1_{\{Y_s \geqslant d\}} ds \leqslant CN^{-1} + \mathbb{E}\int_0^{T\wedge t} \left(\mu(\boldsymbol{X}_s, d) + CZ_s^{d+1} + CN^{-1}\right) ds.$$

Set $g(t) = \mathbb{E}(A_d(T \wedge t))$, then

$$\begin{aligned} g(t) = N\lambda \mathbb{E}\int_0^{T\wedge t} (X_s^d)^n 1_{\{Y_s \geqslant d\}} ds &\leqslant N\lambda 2^n a_d^n \mathbb{E}\int_0^{T\wedge t} 1_{\{Y_s \geqslant d\}} ds \\ &\leqslant Ca_d^n \left(1 + \int_0^t (N\rho^d \mu(a,d) + 1 + g(s))ds\right) \leqslant CN\tilde{a}_{d+1}(1+t) + C\int_0^t g(s)ds. \end{aligned}$$

Here we have used the estimate $\mu(x,d) \leqslant \rho^d \mu(a,d)$ for $x \in U$. The claimed estimate now follows by Gronwall's lemma. $\square$

Fix $R \in (0, \infty)$ and define

$$\tilde{T} = \inf\{t \geqslant 0 : A_d(t) \geqslant RN\tilde{a}_{d+1}\} \wedge T.$$

Lemma 2.8. *There is a constant $C(\lambda, n) < \infty$ such that, for all $t \geqslant 0$ and all N, we have*

$$\mathbb{E}(A_{d+1}(\tilde{T} \wedge t)) \leqslant C(1+t)R^n \tilde{a}_{d+1}^n + CtR^{n+1}N\tilde{a}_{d+1}^{n+1}.$$

Proof. We argue as in the preceding proof, except now taking $f(x,y) = 1_{\{y \geqslant d+1\}}$, for which $\bar{f}(x) = 0$. We obtain

$$\mathbb{E}\int_0^{\tilde{T}\wedge t} 1_{\{Y_s \geqslant d+1\}} ds \leqslant CN^{-1} + C\mathbb{E}\int_0^{\tilde{T}\wedge t} \left(Z_s^{d+1} + CN^{-1}\right) ds$$

and hence

$$\begin{aligned} \mathbb{E}(A_{d+1}(\tilde{T} \wedge t)) &= N\lambda \mathbb{E}\int_0^{\tilde{T}\wedge t} (Z_s^{d+1})^n 1_{\{Y_s \geqslant d+1\}} ds \\ &\leqslant N\lambda (R\tilde{a}_{d+1})^n \mathbb{E}\int_0^{\tilde{T}\wedge t} 1_{\{Y_s \geqslant d+1\}} ds \leqslant C(1+t)R^n \tilde{a}_{d+1}^n + CtR^{n+1}N\tilde{a}_{d+1}^{n+1}. \end{aligned}$$

$\square$

2.10 Proof of Theorem 2.3

Recall that α is defined by (30) and that $\alpha \in (2n, 2n+1)$. Note that $\alpha^2-(2n+1)\alpha+n=0$. Recall that $\kappa=(2\alpha)^{-1}$ and

$$d = d(N) = \sup\{k \in \mathbb{N} : Na_k > N^\kappa\}.$$

The asymptotic growth rate (31) follows from Theorem 2.2. We shall use without further comment below the inequalities

$$C^{-1}a_k^{1/\alpha} \leqslant a_{k+1} \leqslant Ca_k^\alpha, \quad k \geqslant 0$$

proved in Proposition 2.5 and the inequalities

$$a_{d+1} \leqslant N^{-(1-\kappa)} \leqslant a_d, \quad d \leqslant \log\log N,$$

the last being valid for all sufficiently large N.

By the truncation estimate, Lemma 2.6, we have

$$\sup_{t\leqslant t_0}\sup_{k\leqslant d}\frac{|z_k(t)-x_k(t)|}{\sqrt{a_k}} \leqslant t_0\frac{a_{d+1}}{\sqrt{a_d}} \leqslant Ct_0a_{d+1}^{1-1/(2\alpha)} \leqslant Ct_0N^{-1/2}.$$

Since $\phi(N) \to \infty$ as $N \to \infty$ it will therefore suffice to show (32) with $(z(t) : t \geqslant 0)$ replaced by $(x(t) : t \geqslant 0)$.

We apply the general procedure of Subsection 1.3. Take as norm scales $\sigma_k = \sqrt{a_k}$ so that

$$\|x\| = \max_k |x_k|/\sqrt{a_k}, \quad x \in \mathbb{R}^d.$$

We now identify suitable regularity constants $\Lambda, B, \tau, J, J_1(b), J(\mu), K$. We write C for a finite positive constant which may depend on λ and n and whose value may vary from line to line. We shall see that, as $N \to \infty$, the inequalities between these regularity constants required in Theorem 1.6 become valid. The maximum jump rate is bounded above by

$$\Lambda = N(1+\lambda) = CN.$$

We refer to the form of $b(x,y)$ given at (41) and note that, for $x \in U$ and $y \in I$,

$$\|b(x,y)\| \leqslant B = 2^n a_d^{-1/2+n/\alpha} = Ca_d^{-1/2+n/\alpha}.$$

We showed in Subsection 2.5 the following upper bound on the mean coupling time of our coupling mechanism

$$m(x,y,\bar{y}) \leqslant \tau = \frac{4}{N\lambda(1-\lambda)^2} = CN^{-1}.$$

We refer to Subsection 1.3 for the definitions of the jump bounds $J, J_1(b), J(\mu)$ and leave the reader to check the validity of the following inequalities

$$J \leqslant N^{-1} a_d^{-1/2}, \quad J_1(b) \leqslant C N^{-1} a_d^{-1/2+(n-1)/\alpha}, \quad J(\mu) \leqslant 2nN^{-1}.$$

Recall from (46) the form of $\bar{b}$. In estimating the Lipschitz constant K for $\bar{b}$ on U, note first that, for $x \in U$ and for $j = 1, \ldots, k-1$,

$$\left| \frac{\partial}{\partial x_j} x_{k-1}^n \mu(x, k-1) \right| \leqslant C x_{k-1}^n \mu(x, k-1)(x_j^{-1} + 1).$$

Here we have used the explicit form (26) of $\mu(x, k-1)$ and the fact that $(1 - p_{j-1}(x))^{-1} \leqslant C$ on U. Note also the inequalities

$$x_{k-1}^{2n-1} \sqrt{\frac{a_{k-1}}{a_k}} \leqslant 2^{2n-1} a_{k-1}^{2n-1/2-\alpha/2} \leqslant C, \quad \sum_{j=1}^{\infty} \sqrt{a_j} \leqslant C.$$

We find after some further straightforward estimation that we can take $K = C$.

Recall the choice of function ϕ in the statement of Theorem 2.3. Set

$$\varepsilon = \sqrt{\frac{\phi(N)}{N}}, \quad \delta = \varepsilon e^{-Kt_0}/7, \quad \delta(\beta, b) = \delta, \quad \delta(\gamma, g) = \delta/(2\tau B).$$

Recall that $\boldsymbol{X}_0 = (1/N, 0, \ldots, 0)$ and $x_0 = 0$ and that the driving rate ν for the coupling mechanism is equal to $N\lambda$. It is now straightforward to check that all the inequalities required in the statement of Theorem 1.5 are valid, for all sufficiently large N.

We now check the tube condition of Theorem 1.5. The inequalities $0 \leqslant x_d(\xi) \leqslant \ldots \leqslant x_1(\xi) \leqslant 1$ hold for all $\xi \in S$. By a monotonicity property established in the proof of Theorem 2.1, we have $x_k(t) \leqslant a_k$ for all $t \geqslant 0$ and for $k = 1, \ldots, d$. Hence, for N sufficiently large, if $\|\boldsymbol{x}(\xi) - x(t)\| \leqslant 2\varepsilon$ for some $t \geqslant 0$, then $x^k(\xi) \leqslant a_k + 2\varepsilon\sqrt{a_k} \leqslant 2a_k$ and $x^1(\xi) \leqslant a_1 + 2\varepsilon\sqrt{a_1} \leqslant \lambda + (1-\lambda)/2 \leqslant (1+\lambda)/2$, so $\boldsymbol{x}(\xi) \in U$ and the tube condition is satisfied.

We turn now to the extra conditions needed to apply Theorem 1.6. We noted in Subsection 2.8 the quadratic variation bounds

$$a_k(y) \leqslant A\sigma_k^2, \quad \bar{a}_k(x) \leqslant \bar{A}\sigma_k^2$$

valid for all $x \in U$ and $y \in I$, where

$$A = C/(N a_d^{1-n/\alpha}), \quad \bar{A} = C(\log N)^C/N$$

and where $\bar{A} \leqslant A \leqslant \Lambda \bar{J}^2$ for sufficiently large N. It is now straightforward to check, also for N sufficiently large, that the remaining inequalities required in the statement of Theorem 1.6 hold. Theorem 1.6 therefore applies to give

$$\mathbb{P}\left(\sup_{t \leqslant t_0} \|\boldsymbol{X}_t - x_t\| > \varepsilon \right) \leqslant 2de^{-\delta^2/(4\bar{A}t_0)} + 2de^{-(\bar{A}/A)^2 t_0/(6400\Lambda\tau^2)} + \mathbb{P}(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c) \quad (50)$$

Now, for N sufficiently large, we have $d \leqslant \log\log N$ and, by our choice of ϕ and κ,

$$\delta^2/(4\bar{A}t_0) \geqslant \phi(N)/((\log N)^C t_0) \geqslant \log N$$

and

$$(\bar{A}/A)^2 t_0/(6400\Lambda\tau^2) \geqslant \log N.$$

Hence the first and second terms on the right hand side of (50) tend to 0 as $N \to \infty$.

Recall from (16) and (17) the form of the events $\Omega(\beta, b)$ and $\Omega(\gamma, g)$. In the present example, the complementary exceptional events arise either as a result of truncation or because of finite N effects in the fast variable dynamics, as shown by the equation (42) and the estimate (44). Recall that $\delta(\beta, b) = \delta$ and $\delta(\gamma, g) = \delta/(2\tau B)$. Then

$$\Omega(\beta, b)^c \subseteq \left\{ \int_0^{T\wedge t_0} \frac{Z_t^{d+1}}{\sqrt{a_d}} dt \geqslant \delta(\beta, b) \right\} \subseteq \left\{ A_d(T \wedge t_0) \geqslant \frac{N\delta\sqrt{a_d}}{t_0} \right\}. \tag{51}$$

It is straightforward to check that, for all sufficiently large N, $\delta(\gamma, g) \geqslant 2t_0(1 + \lambda n)$, which implies that

$$\Omega(\gamma, g)^c \subseteq \left\{ \int_0^{T\wedge t_0} (1 + \lambda n + N\lambda n Z_t^{d+1}) dt \geqslant \delta(\gamma, g) \right\} \subseteq \left\{ A_d(T \wedge t_0) \geqslant \frac{\delta}{4\lambda n t_0 \tau B} \right\}. \tag{52}$$

To see that $\mathbb{P}(\Omega(\beta, b)^c \cup \Omega(\gamma, g)^c) \to 0$ as $N \to \infty$ we use the bound on $\mathbb{E}(A_d(T \wedge t_0))$ proved in Lemma 2.7 and Markov's inequality. It then suffices to show that in the limit $N \to \infty$,

$$C(a_d^n + \rho^d N a_{d+1}) e^{Ct_0} \ll \frac{\sqrt{\phi(N) N a_d}}{4nt_0}.$$

For the term involving a_d^n this is easy. For the other term, involving Na_{d+1}, we can check that in fact

$$Na_{d+1} \ll \sqrt{Na_d}, \quad \rho^d \leqslant (\log N)^C \ll \sqrt{\phi(N)}.$$

This completes the proof of (32). The limit (33) follows immediately from Lemma 2.7 using Markov's inequality. Finally note that, as $N \to \infty$,

$$\tilde{a}_{d+1} \leqslant CN^{-1} a_d^n + (\log N)^C a_{d+1} \leqslant C(N^{-1} + (\log N)^C N^{-1+\kappa}) \to 0$$

and

$$N\tilde{a}_{d+1}^{n+1} \leqslant C(N^{-(1-\kappa)(n+1)/\alpha} + (\log N)^C N^{1-(1-\kappa)(n+1)}) \to 0.$$

Then the limit (34) follows from (33) and Lemma 2.8 using Markov's inequality. $\square$

2.11 Monotonicity of the queueing model

We prove here a natural monotonicity property of the supermarket model with memory which is a microscopic counterpart of the monotonicity of solutions to the differential equation (28) shown in Theorem 2.1. We do not rely on this result in the rest of the paper.

First we construct, on a single probability space, for all $\xi = (z, y) \in S$, a version $X = X(\xi)$ of the supermarket model with memory starting from ξ. Set $y_1 = y_1(\xi) = y$ and determine $y_i = y_i(\xi) \in \mathbb{Z}^+$ for $i = 2, \dots, N$ by the conditions

$$y_2 \leqslant \dots \leqslant y_N, \quad z_k = |\{i \in \{1, \dots, N\} : y_i \geqslant k\}|/N, \quad k \in \mathbb{N}.$$

We work on the canonical space of a marked Poisson process of rate $N(1+\lambda)$, where the marks are either, with probability $1/(1+\lambda)$, independent copies of a uniform random variable J in $\{1, \dots, N\}$ or, with probability $\lambda/(1+\lambda)$, independent copies of a uniform random sample $(J_1, \dots, J_n)$ from $\{1, \dots, N\}$. Fix $\xi = (z, y) \in S$ and define a process $X = X(\xi) = (X_t : t \geqslant 0)$ in S as follows. Set $X_t = \xi$ for all $t < T$, where T is the first jump time of the Poisson process. If the first mark is a random variable, J say, take the sequence $y_1, \dots, y_N$ and replace y_J by $(y_J - 1)^+$ to obtain a sequence $u_1, \dots, u_N$ say; set $\tilde{y}_1 = u_1$ and write $u_2, \dots, u_N$ in non-decreasing order to obtain $\tilde{y}_2 \leqslant \dots \leqslant \tilde{y}_N$. If the first mark is a random sample, $(J_1, \dots, J_n)$ say, select components $(y_i : i \in \{1, J_1, \dots, J_n\})$ and write these in non-decreasing order, $w_1 \leqslant \dots \leqslant w_m$ say; replace w_1 by $w_1 + 1$ and write the resulting sequence again in non-decreasing order, $v_1 \leqslant \dots \leqslant v_m$ say; set $\tilde{y}_1 = v_1$ and write $v_2, \dots, v_m$ combined with the unselected components $(y_i : i \notin \{1, J_1, \dots, J_n\})$ in non-decreasing order to obtain $\tilde{y}_2 \leqslant \dots \leqslant \tilde{y}_N$. Set $X_T = ((Z_T^k : k \in \mathbb{N}), Y_T)$, where

$$Z_T^k = |\{i \in \{1, \dots, N\} : \tilde{y}_i \geqslant k\}|/N, \quad k \in \mathbb{N}, \quad Y_T = \tilde{y}_1$$

and repeat the construction from X_T in the usual way.

For $\xi, \xi' \in S$ write $\xi \leqslant \xi'$ if $y_i(\xi) \leqslant y_i(\xi')$ for $i = 1, \dots, N$.

Theorem 2.9. *Let $\xi, \xi' \in S$ with $\xi \leqslant \xi'$. Then $X_t(\xi) \leqslant X_t(\xi')$ for all $t \geqslant 0$.*

Proof. It will suffice to check that the desired inequality holds at the first jump time T, that is to say, with obvious notation, that $\tilde{y}_i \leqslant \tilde{y}_i'$ for all i. Note that if $a_i \leqslant b_i$ for all i for two sequences $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$ then the same is true for their non-decreasing rearrangements. In the case where the first mark is a random variable J, since $y_i \leqslant y_i'$ for all i, we have $u_i \leqslant u_i'$ for all i and so $\tilde{y}_i \leqslant \tilde{y}_i'$ for all i. On the other hand, when the first mark is a random sample $(J_1, \dots, J_n)$, we have $w_j \leqslant w_j'$ for all j, so $v_j \leqslant v_j'$ for all j, and so $\tilde{y}_i \leqslant \tilde{y}_i'$ for all i. □

References

[1] Karen Ball, Thomas G. Kurtz, Lea Popovic, and Greg Rempala. Asymptotic analysis of multiscale approximations to reaction networks. *Ann. Appl. Probab.*, 16(4):1925–1961, 2006.

[2] R. W. R. Darling and J. R. Norris. Differential equation approximations for Markov chains. *Probab. Surv.*, 5:37–79, 2008.

[3] Carl Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *J. Appl. Probab.*, 37:198–210, 2000.

[4] Carl Graham. Functional central limit theorems for a large network in which customers join the shortest of several queues. *PTRF*, 131:97–120, 2004.

[5] Malwina Luczak and Colin McDiarmid. On the maximum queue length in the supermarket model. *Ann. Probab.*, 34:493–527, 2006.

[6] Malwina Luczak and Colin McDiarmid. Asymptotic distributions and chaos for the supermarket model. *Elec. J. Probab.*, 12:75–99, 2007.

[7] Malwina J. Luczak and James Norris. Strong approximation for the supermarket model. *Ann. Appl. Probab.*, 15(3):2038–2061, 2005.

[8] Prabhakar Balaji Mitzenmacher, Michael and Devavrat Shah. Load balancing with memory. *Proc. 43rd Ann. IEEE Symp. Found. Comp. Sci.*, 2002.

[9] Dobrushin R.L. Vvedenskaya, N. and F.I. Karpelevich. Queueing system with selection of the shortest of two queues. *Prob. Inf. Transm.*, 32:15–27, 1996.